# *k*NN

and the bias-variance tradeoff

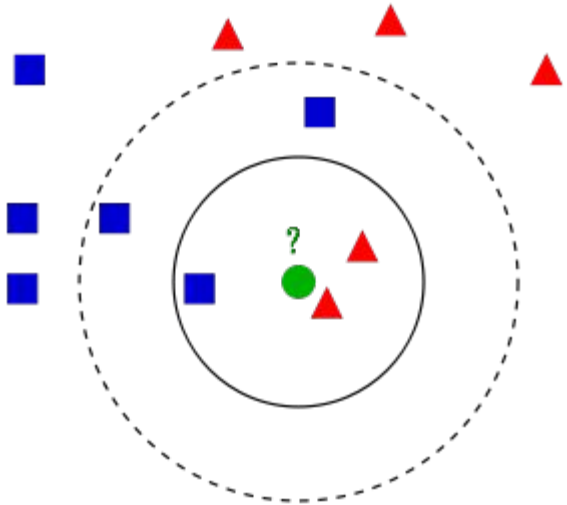



Image Source

# iris

3 species (i.e., classes) of iris

- Iris setosa
- Iris versicolor
- Iris virginica

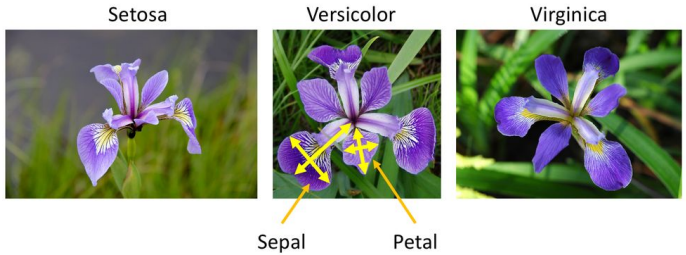


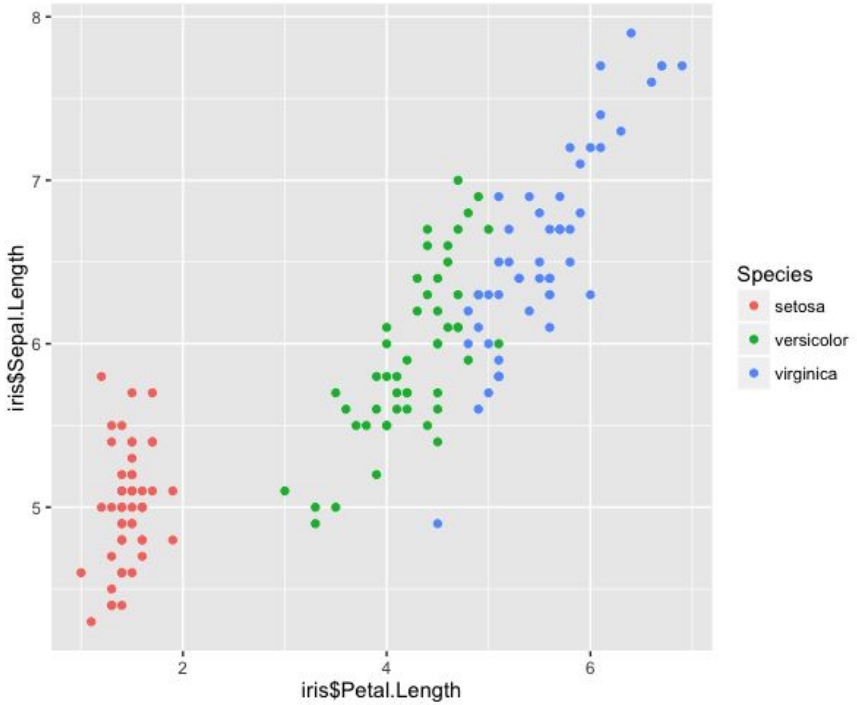

Image Source

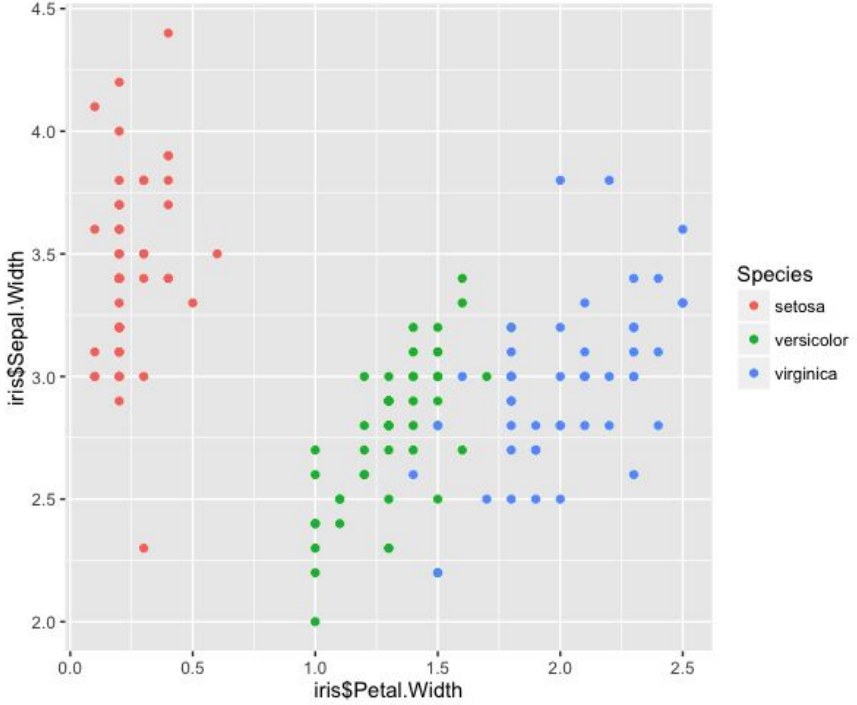
# iris

- 50 observations per species
- 4 variables per observation
  - Sepal length
  - Sepal width
  - Petal length
  - Petal width

| Id | SepalLengthCm | SepalWidthCm | PetalLengthCm | PetalWidthCm | Species |
|---|---|---|---|---|---|
| 1 | 5.1 | 3.5 | 1.4 | 0.2 | Iris-setosa |
| 2 | 4.9 | 3 | 1.4 | 0.2 | Iris-setosa |
| 3 | 4.7 | 3.2 | 1.3 | 0.2 | Iris-setosa |
| 4 | 4.6 | 3.1 | 1.5 | 0.2 | Iris-setosa |
| 5 | 5 | 3.6 | 1.4 | 0.2 | Iris-setosa |
| 6 | 5.4 | 3.9 | 1.7 | 0.4 | Iris-setosa |
| 7 | 4.6 | 3.4 | 1.4 | 0.3 | Iris-setosa |
| 8 | 5 | 3.4 | 1.5 | 0.2 | Iris-setosa |
| 9 | 4.4 | 2.9 | 1.4 | 0.2 | Iris-setosa |
| 10 | 4.9 | 3.1 | 1.5 | 0.1 | Iris-setosa |

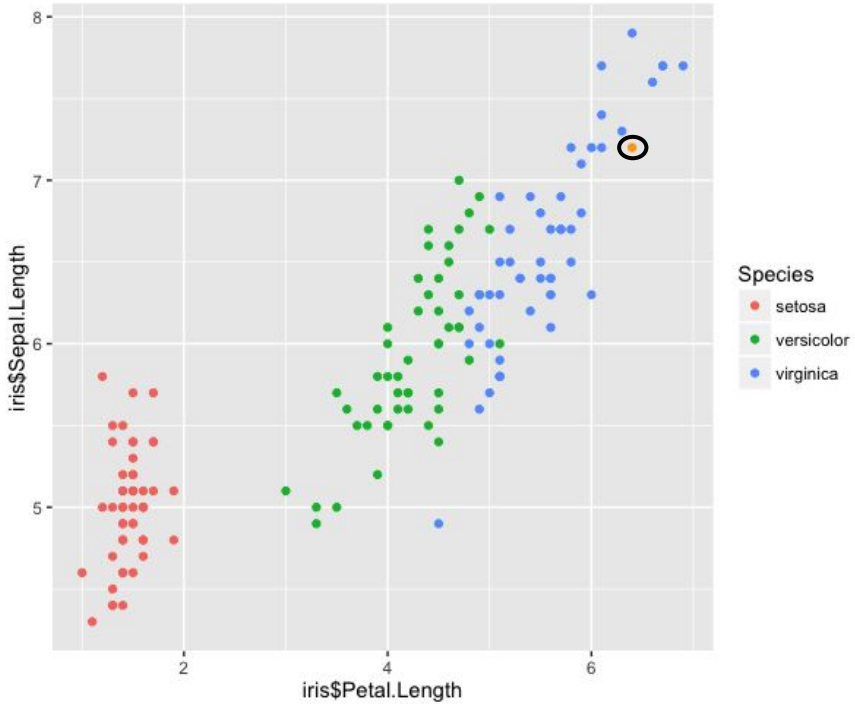
# Visualizing the data



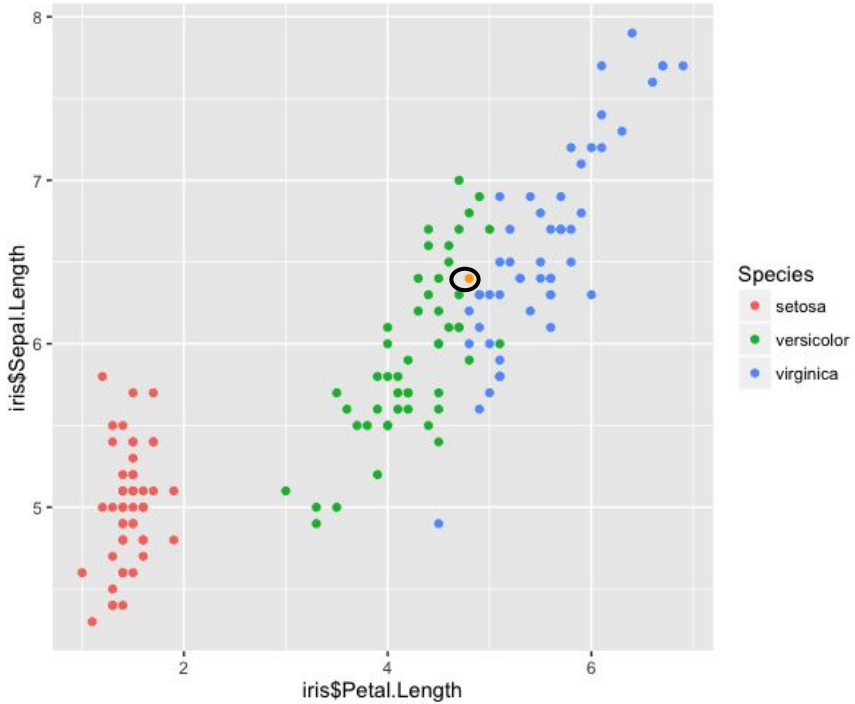
# A new observation

```
new_point <- data.frame
    (Sepal.Length = 7.2,
     Sepal.Width  = 3.2,
     Petal.Length = 6.4,
     Petal.Width  = 2.4)
```



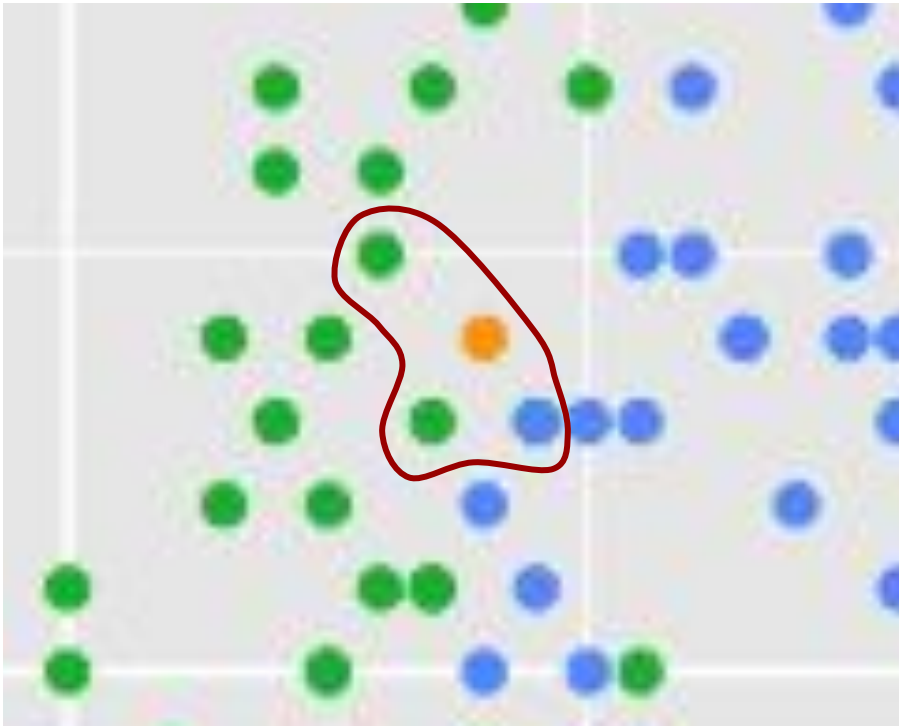
# Another observation

```
new_point <- data.frame
    (Sepal.Length = 6.4,
     Sepal.Width  = 2.8,
     Petal.Length = 4.9,
     Petal.Width  = 1.3)
```



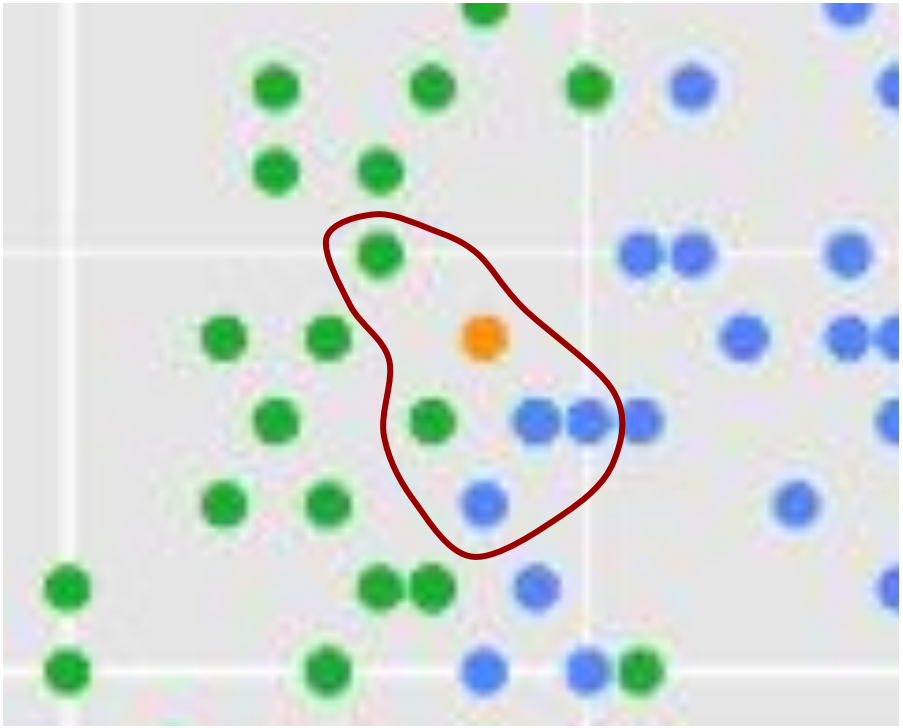
# Another observation

```
[1] k = 3
[1] versicolor
```



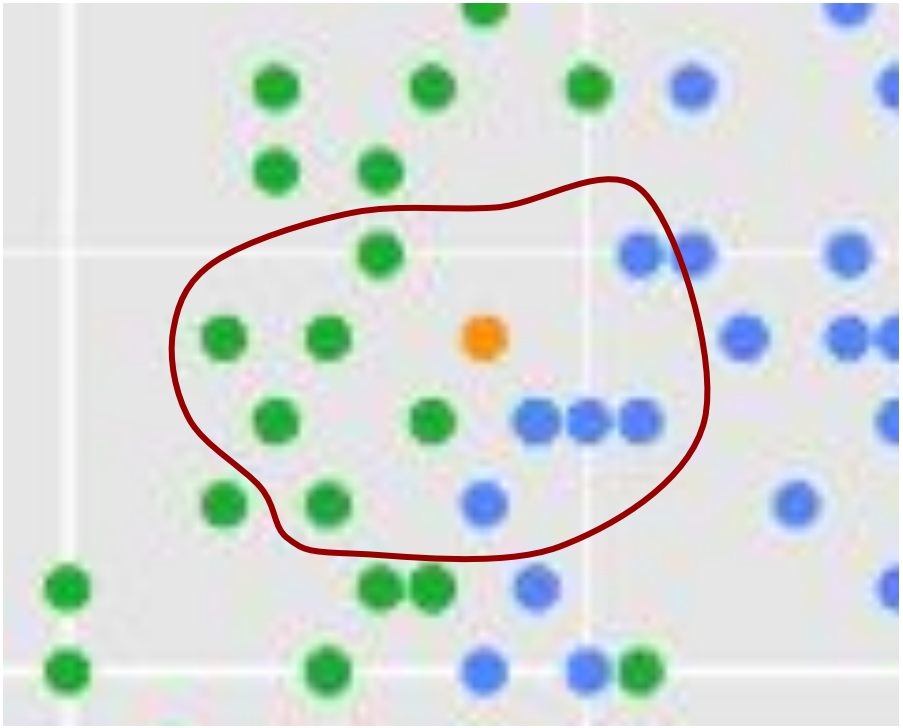
# Another observation

```
[1] k = 3
[1] versicolor
[1] k = 5
[1] virginica
```

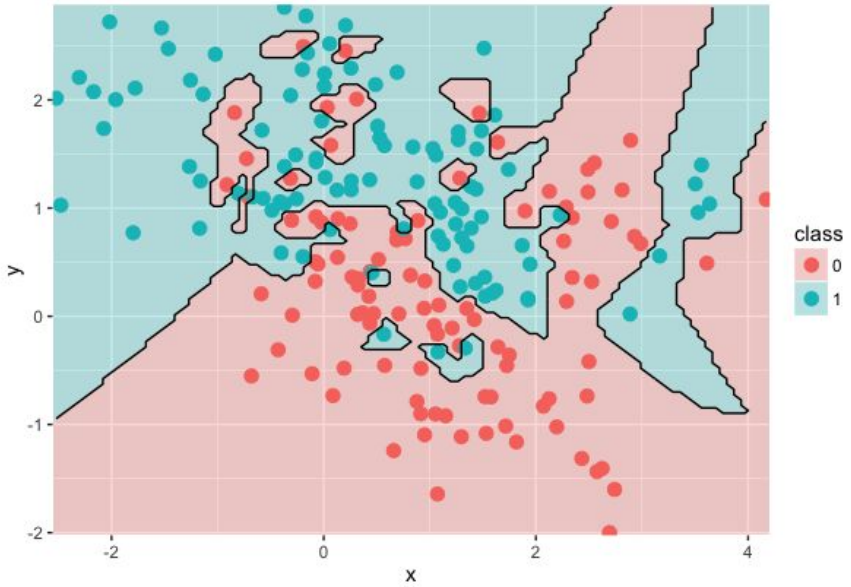


# Another observation

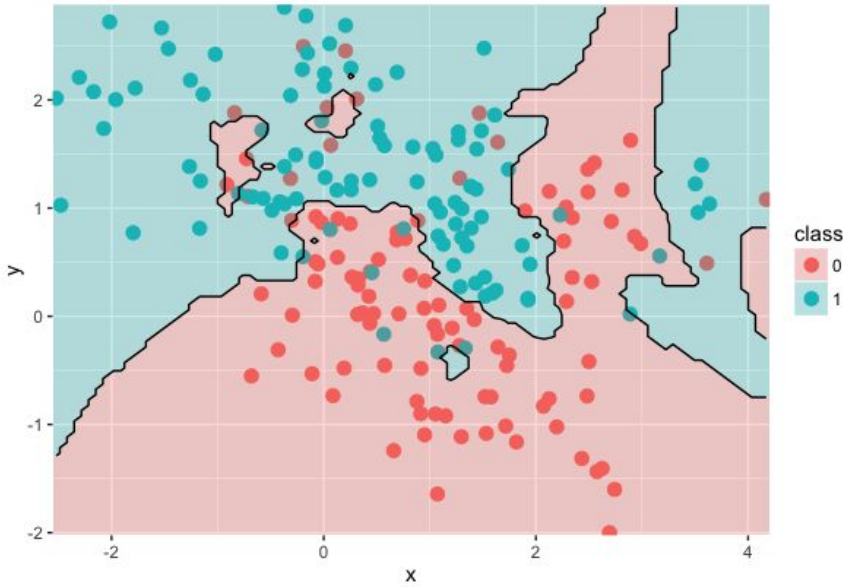
```
[1] k = 3
[1] versicolor
[1] k = 5
[1] virginica
[1] k = 11
[1] versicolor
```



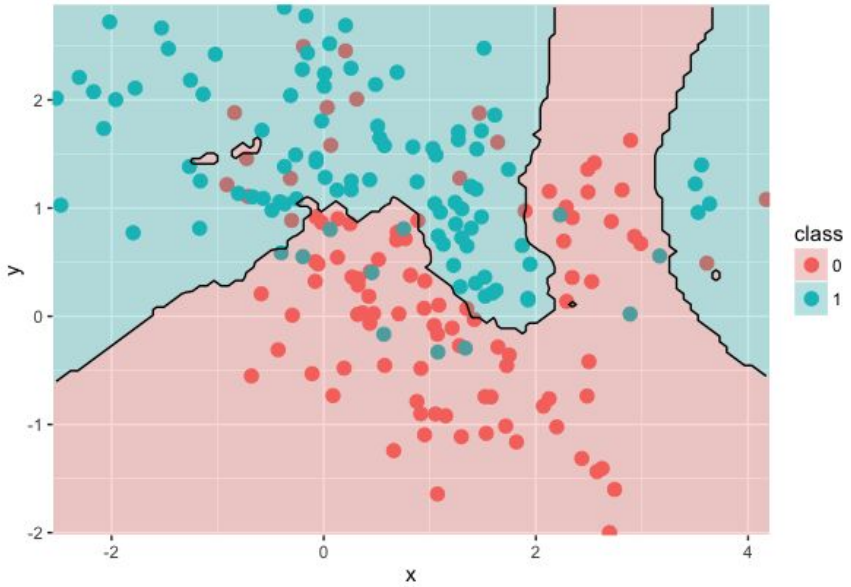
$k = 1$



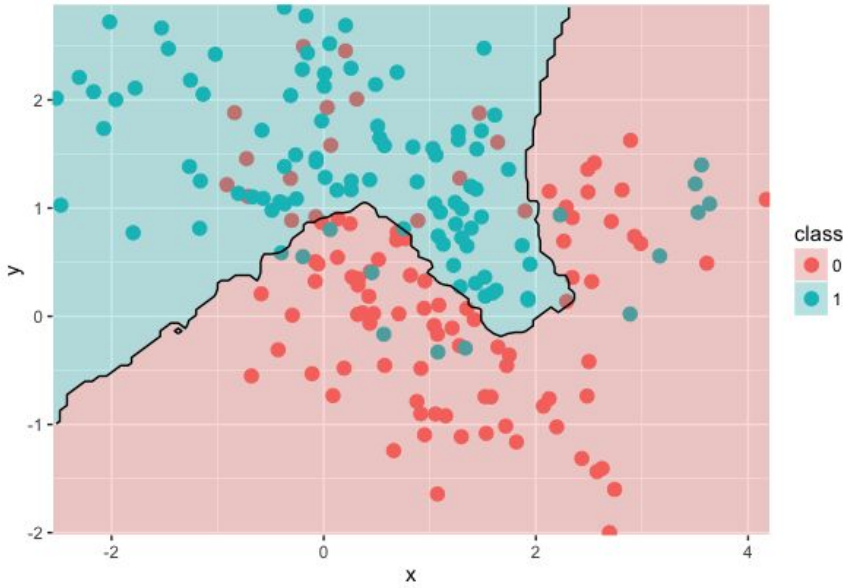
$k = 3$



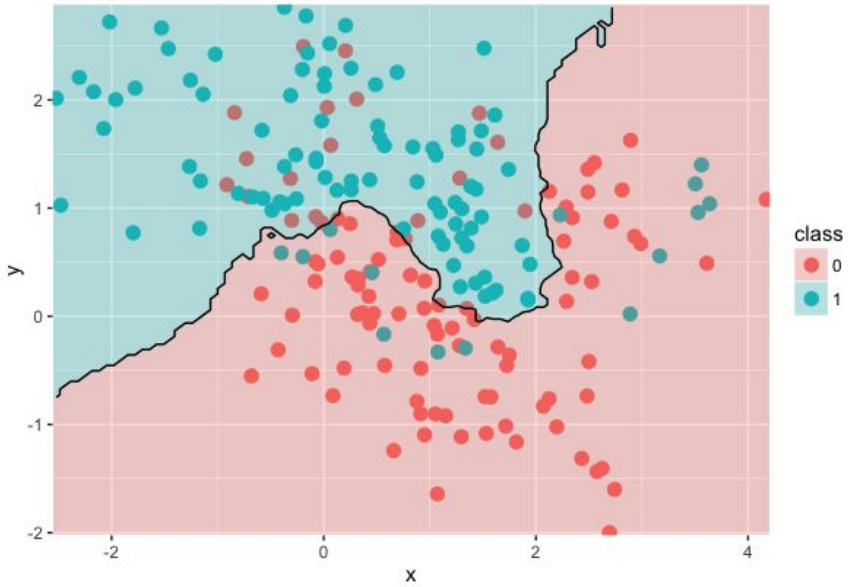
$k = 7$



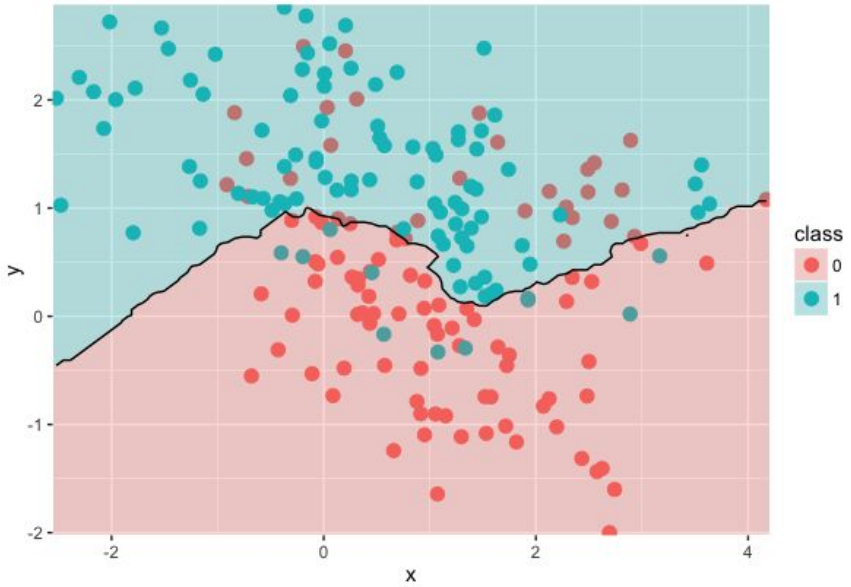
$k = 15$



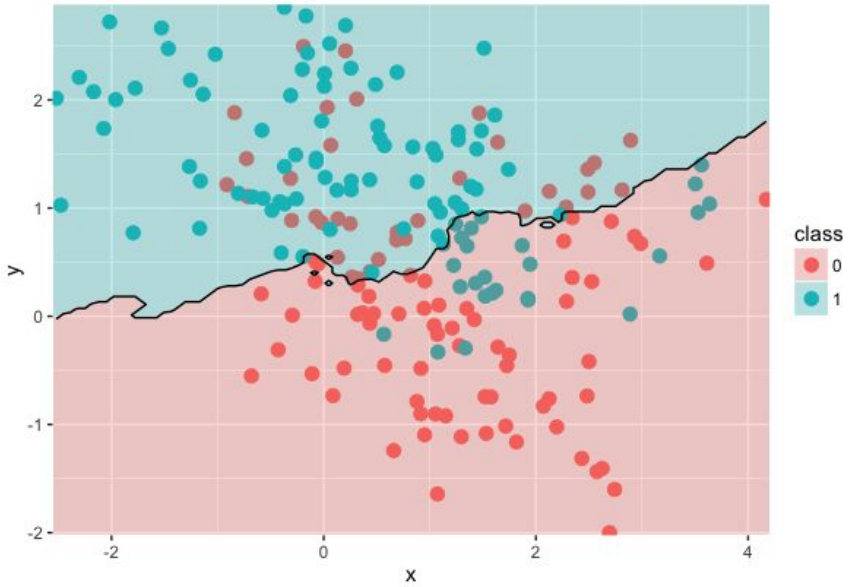
$k = 25$



$k = 51$

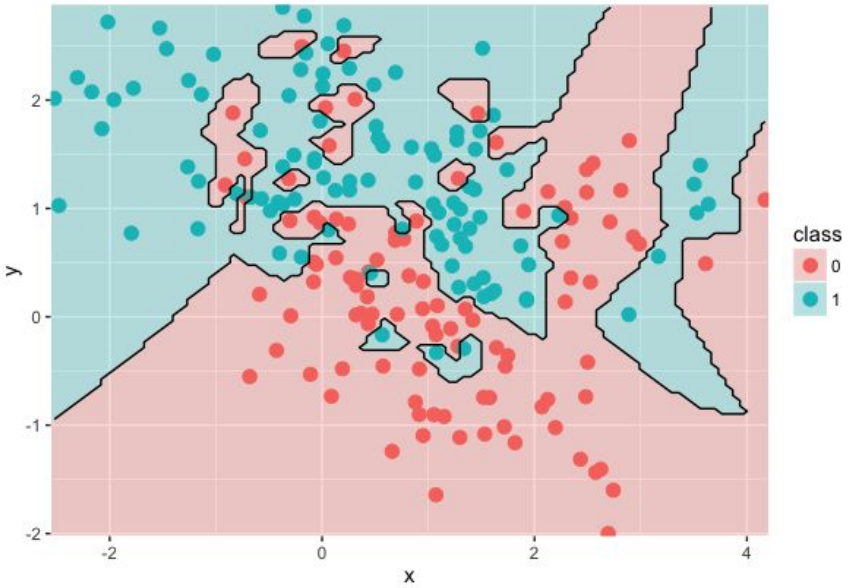


$k = 101$



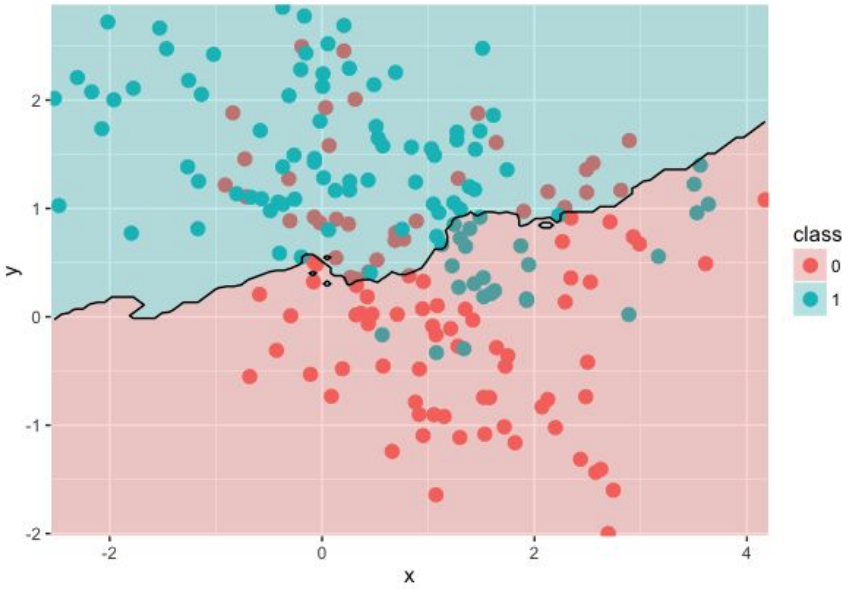
# Small *k*

- $k = 1$
- Low bias
- High variance:
  model varies greatly with the data
- Models like this are overfit
  The model reads too much into the data, extrapolating based on things that aren't necessarily relevant

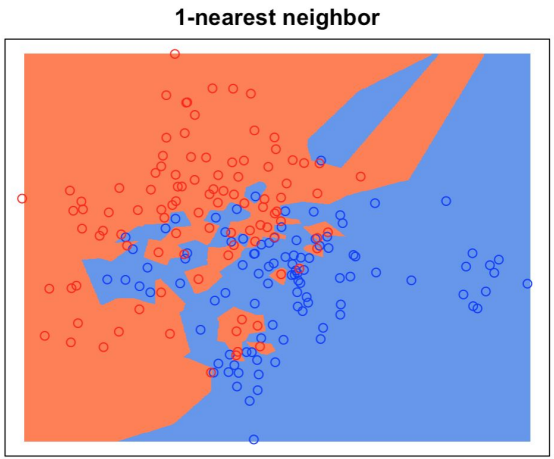


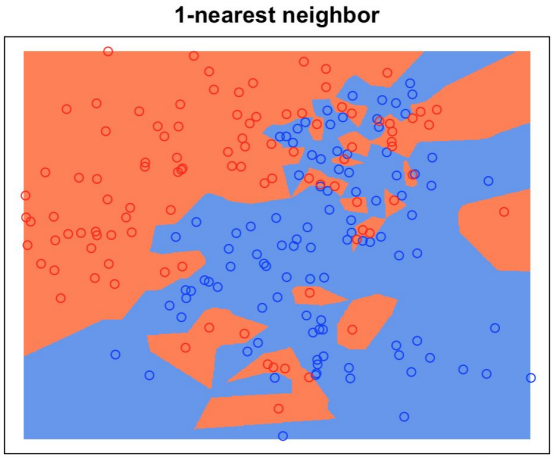
# Large *k*

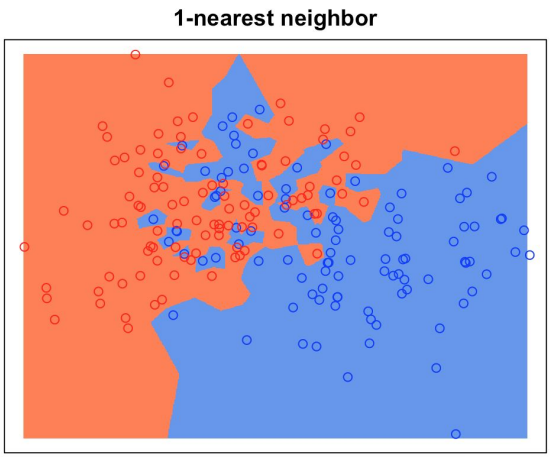
- $k = 101$
- High bias
- Low variance:
  model barely varies with the data
- Rather than being overfit,
  this model is underfit
  The decision boundary doesn't capture enough of the relevant information encoded in the data



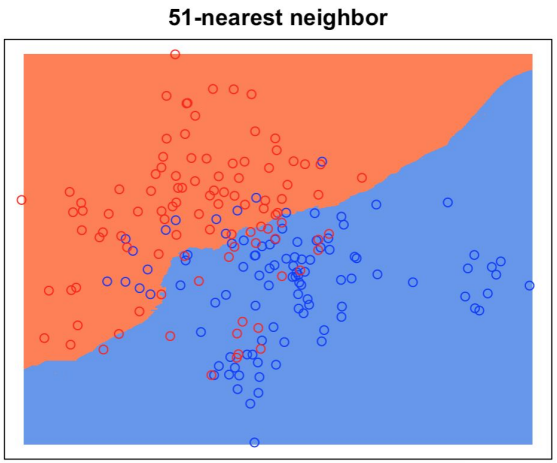
# $k = 1$

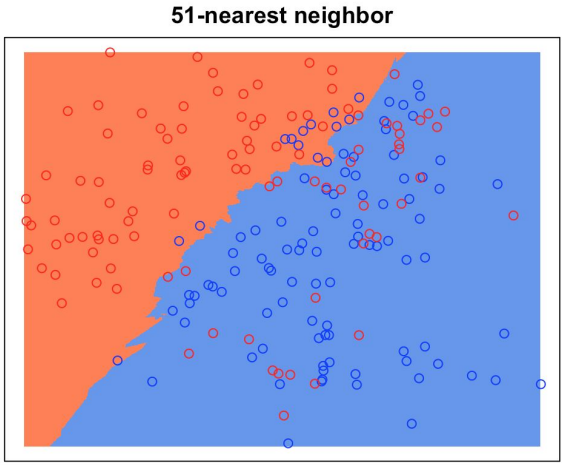


# $k = 51$


51-nearest neighbor

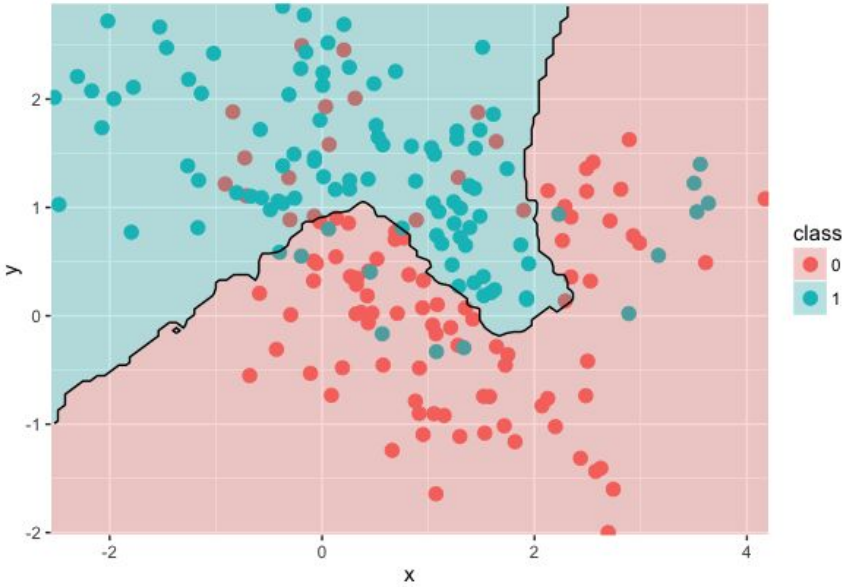

51-nearest neighbor


51-nearest neighbor

# Model Selection

Find a model that balances the bias-variance tradeoff.

- A high value of *k* correspond to a high degree of bias, but contains the variance
- With low values of *k*, the jagged decision boundaries are a sign of high variance
- *k* = 15 seems “just right”



# Key Design Decisions

- Choose *k*
  - Choose a threshold
- Define “neighbor”
  - Define a measure of distance/closeness
    (Make sure measurement values are comparable)
- Decide how to classify based on neighbors’ labels
  - By a majority vote, or
  - By a weighted majority vote (weighted by distance), or …

# *k*-NN caveats

- *k*-NN can be very slow, especially for very large data sets
  - *k*-NN is not a learning algorithm in the traditional sense, because it doesn't actually do any learning: i.e., it doesn't preprocess the data
  - Instead, when it is given a new observation, it calculates the distance between that observation and every existing observation in the data set
- *k*-NN works better with quantitative data than categorical data
  - Data must be quantitative to calculate distances
  - So categorical data must be transformed
- Without clusters in the training data, *k*-NN cannot work well

# Bias-Variance Tradeoff

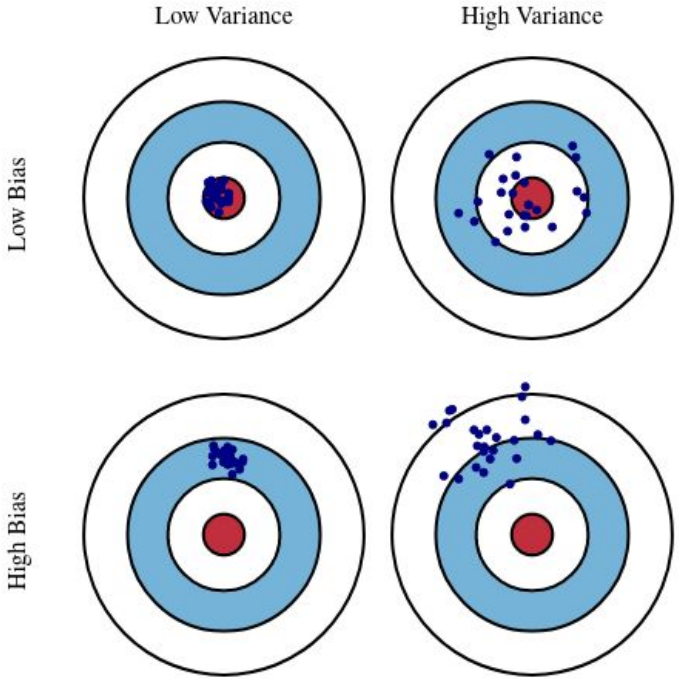



Image Source